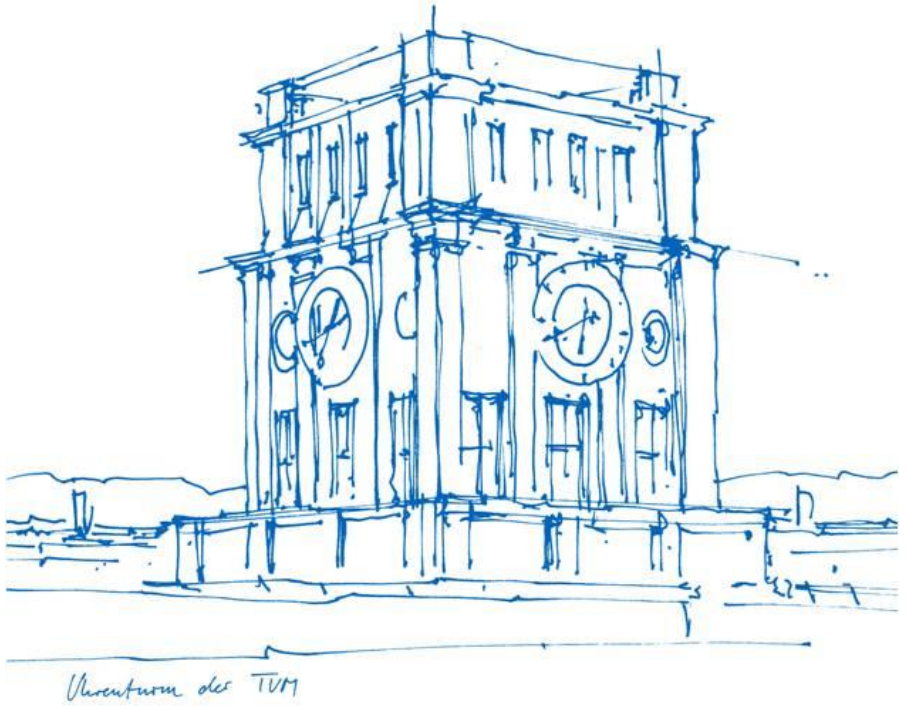
Hardware-Efficient Neural Networks for Low-Latency Multiplexed Superconducting Qubit Readout

Xiaorang Guo and Martin Schulz

Chair of Computer Architecture and Parallel Systems

Technical University of Munich (TUM)

February 23, 2026



Outline

- Background & Motivation
- Methodology – Neural Network Development
- FPGA-based Implementation
- Evaluation
- Conclusion & Future Work

Background – Superconducting Qubit Readout

Key Component of the SC Quantum Computer

- Long Latency
  - Operation cycle
  - Instructions within coherence time
- Error-prone
  - High-fidelity quantum computers

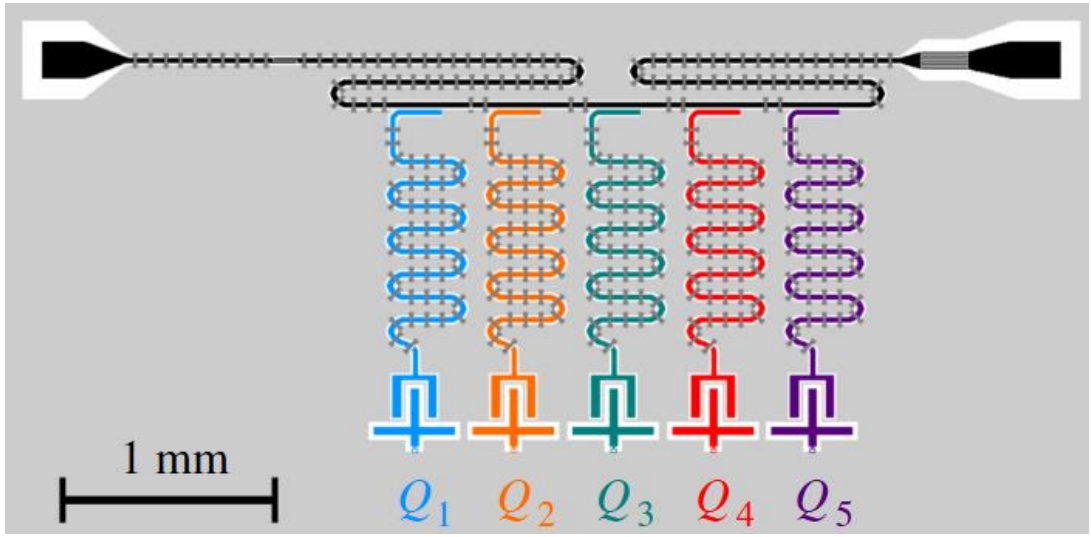
Background – Superconducting Qubit Readout

Key Component of the SC Quantum Computer

- Long Latency
  - Operation cycle
  - Instructions within coherence time
- Error-prone
  - High-fidelity quantum computers

Three main approaches

- Each qubit by a separate resonator
- Several qubits coupled to a single readout resonator
- Frequency-multiplexed readout signals from multiple readout resonators

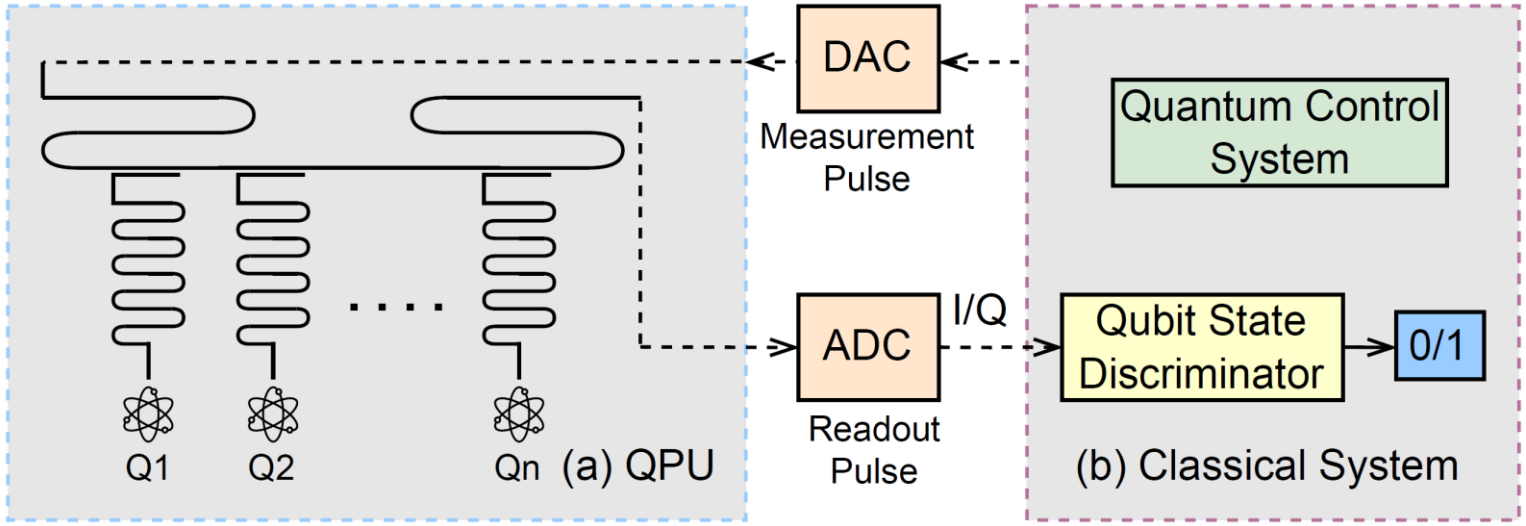


Source: https://doi.org/10.1103/PhysRevApplied.17.014024

Background – Superconducting Qubit Readout

Multi-stage Process

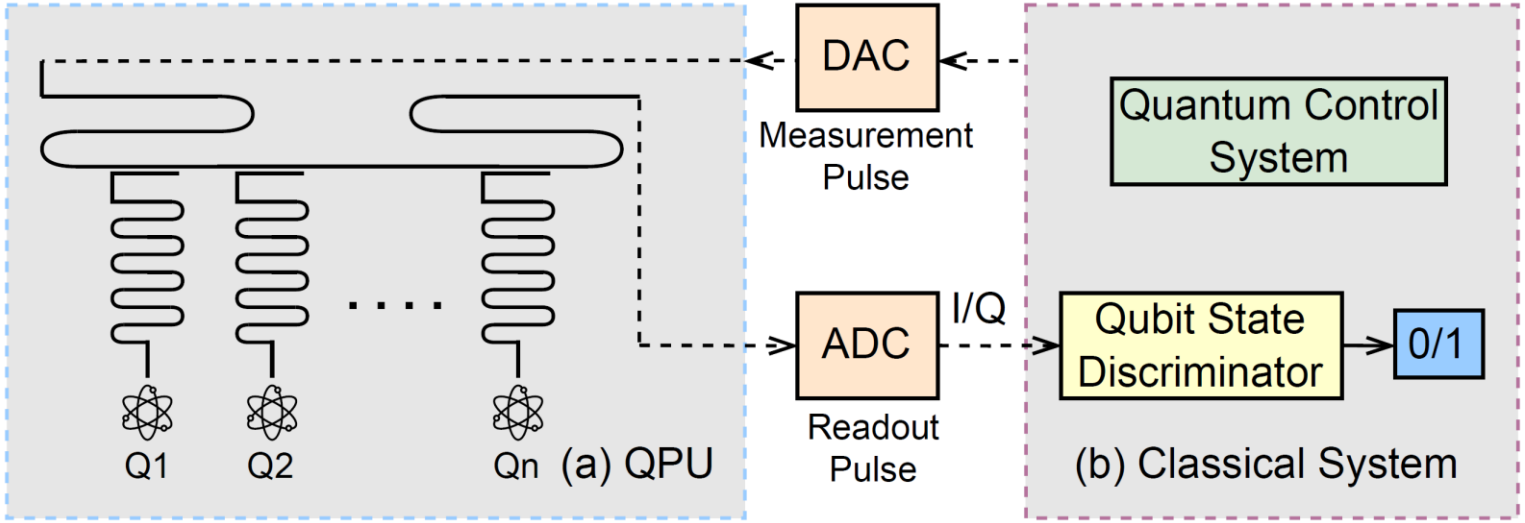
- Analog Microwaves from QPU
- Processes by Analog Digital Converter (ADC)
- Digital Signals Discrimination
  - In-phase (I) and quadrature (Q) signals
  - Single-shot qubit-state discrimination

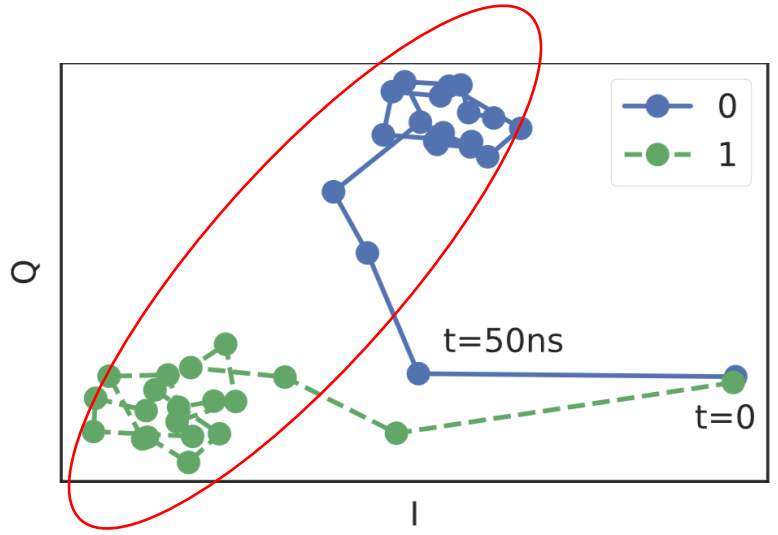


Background – Superconducting Qubit Readout

Multi-stage Process

- Analog Microwaves from QPU
- Processes by Analog Digital Converter (ADC)
- Digital Signals Discrimination
  - In-phase (I) and quadrature (Q) signals
  - Single-shot qubit-state discrimination



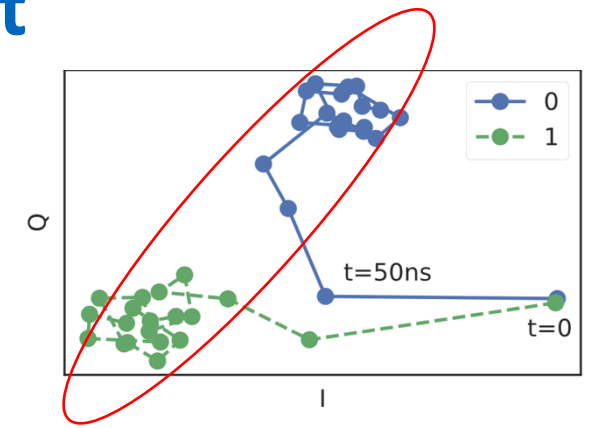


Source: https://dl.acm.org/doi/pdf/10.1145/3579371.3589042

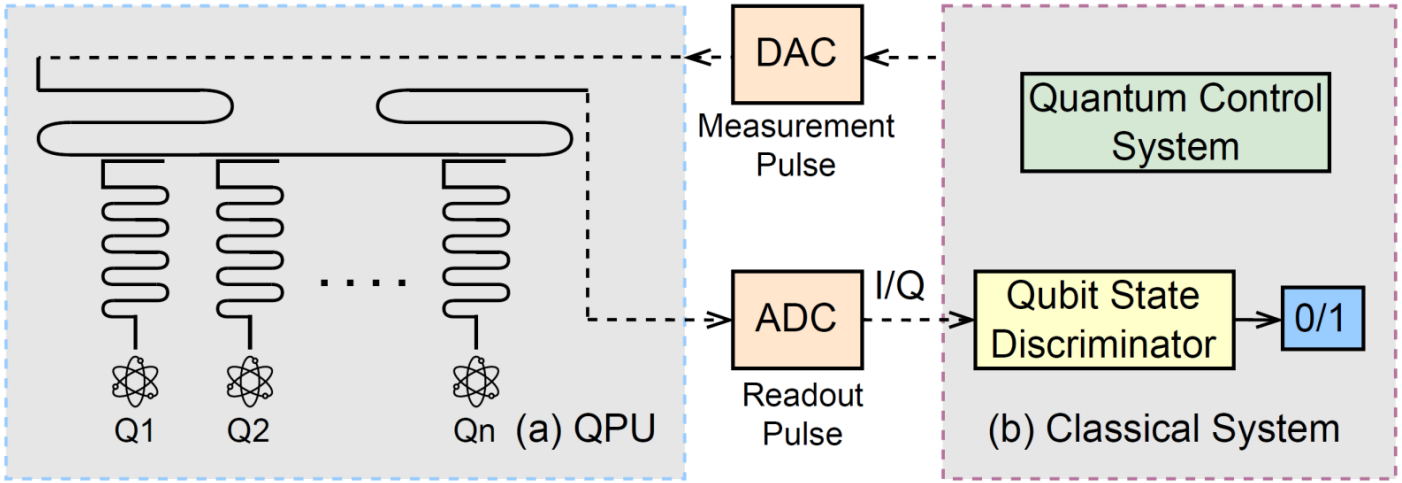
Background – Superconducting Qubit Readout

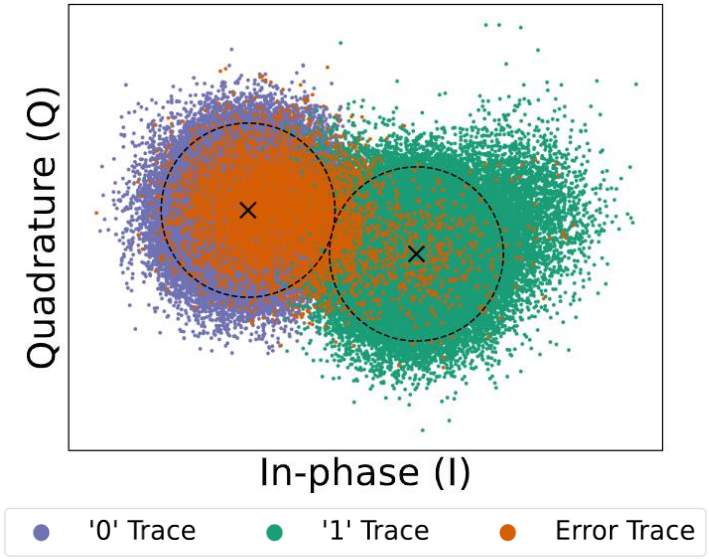
Multi-stage Process

- Analog Microwaves from QPU
- Processes by Analog Digital Converter (ADC)
- Digital Signals Discrimination
  - In-phase (I) and quadrature (Q) signals
  - Single-shot qubit-state discrimination



Source: https://dl.acm.org/doi/pdf/10.1145/3579371.3589042





Related Works

Classical Methods

- Boxcar filters
- Matched filter (MF) windows
- Support vector machines (SVM)

Related Works

Classical Methods

- Boxcar filters
- Matched filter (MF) windows
- Support vector machines (SVM)

Neural Networks

- Convolutional neural networks (CNN)
- Long short-term memory recurrent neural networks (LSTM)
- Fully-connected feedforward neural network (FNN)
  - Best candidate so far

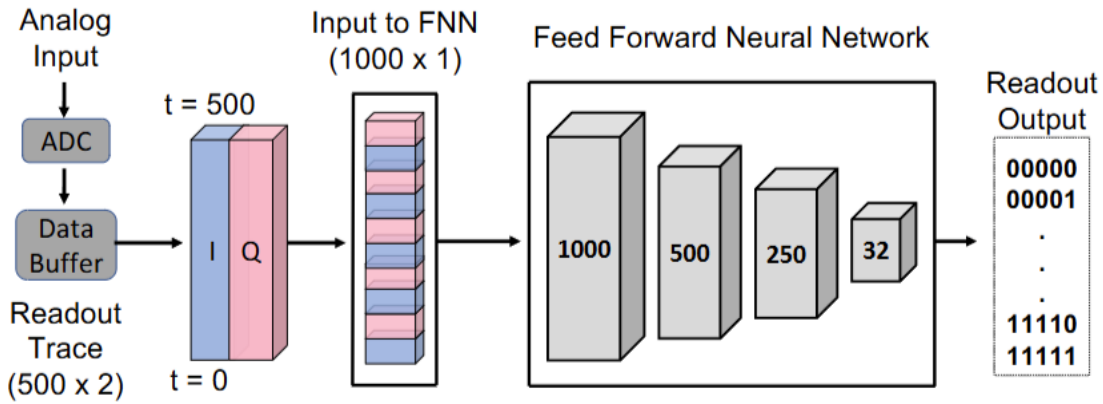
Related Works

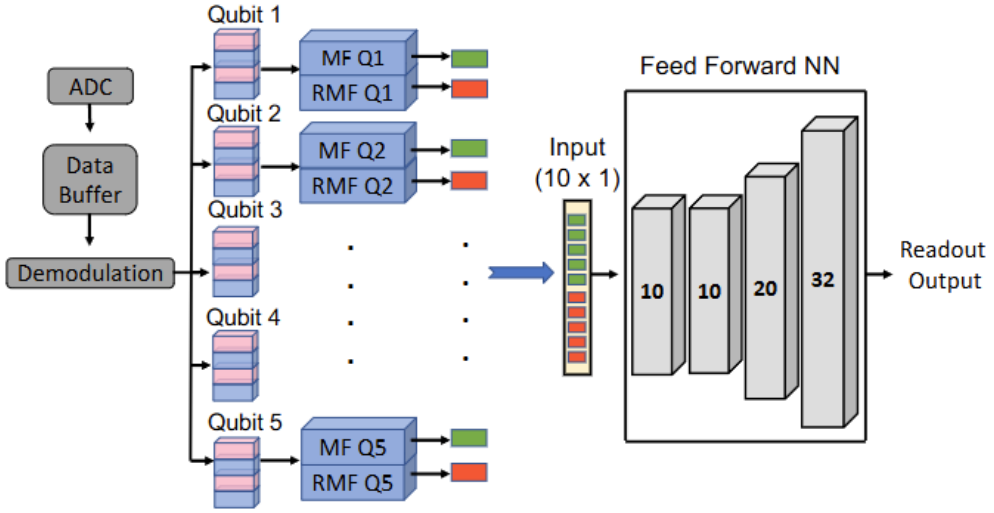
Classical Methods

- Boxcar filters
- Matched filter (MF) windows
- Support vector machines (SVM)

Neural Networks

- Convolutional neural networks (CNN)
- Long short-term memory recurrent neural networks (LSTM)
- Fully-connected feedforward neural network (FNN)
  - Best candidate so far



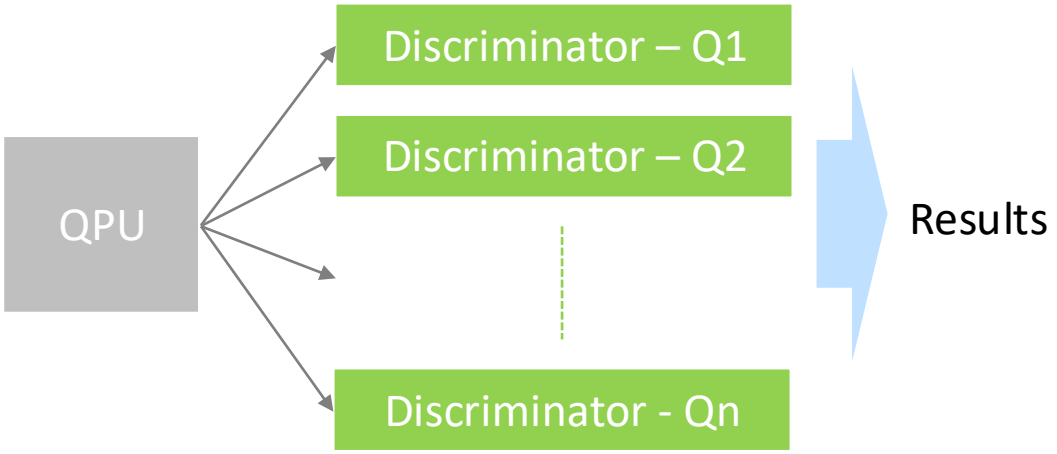


Source: https://dl.acm.org/doi/pdf/10.1145/3579371.3589042

Challenges on the Optimal Way

Mid-circuit Measurement

- Current architecture mostly requires simultaneous readout of all qubits
- Hindering the applicability in mid-circuit measurement scenarios
  - Individual readout is required



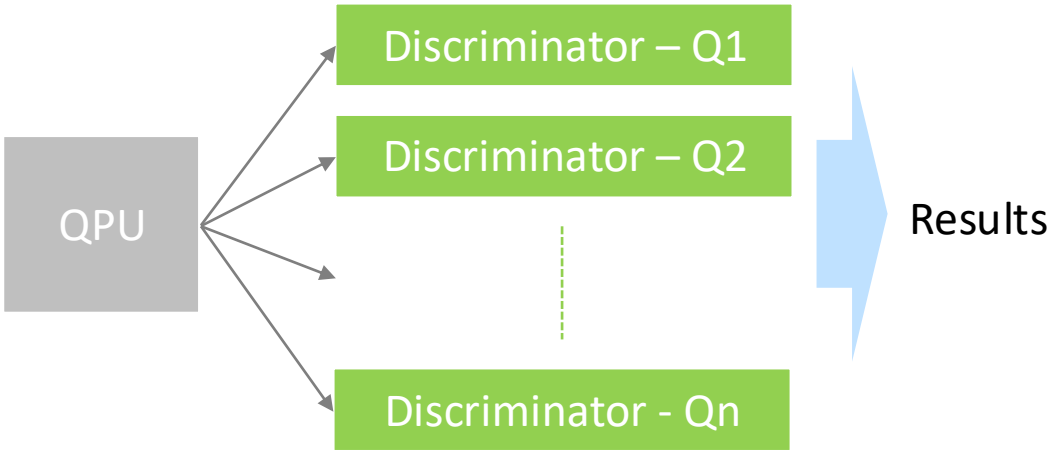
Challenges on the Optimal Way

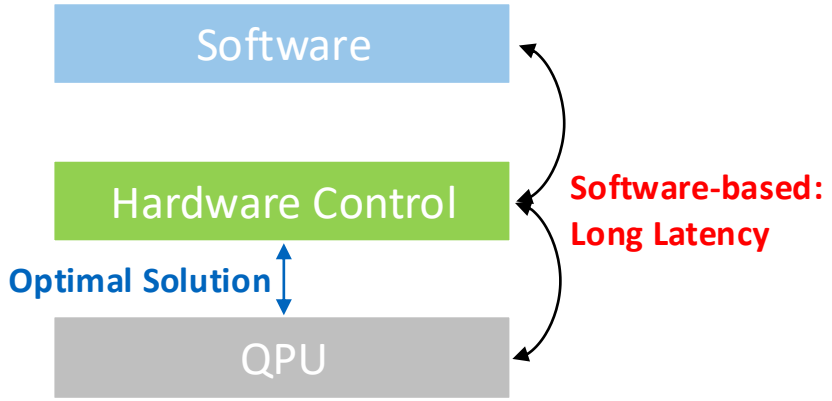
Mid-circuit Measurement

- Current architecture mostly requires simultaneous readout of all qubits
- Hindering the applicability in mid-circuit measurement scenarios
  - Individual readout is required

Migration to FPGAs

- Software-based methods -> Long process & communication latency
- Existing designs are too heavy to scale to large qubit numbers on the hardware

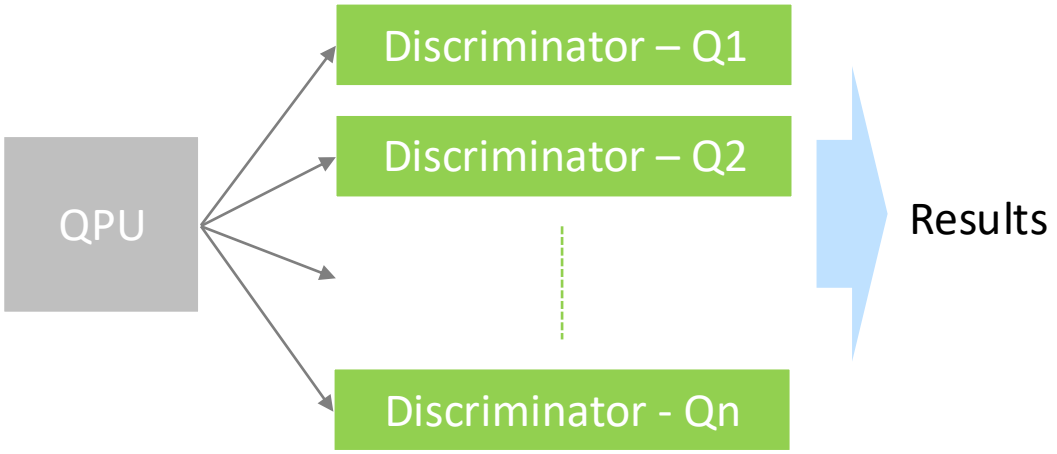




Challenges on the Optimal Way

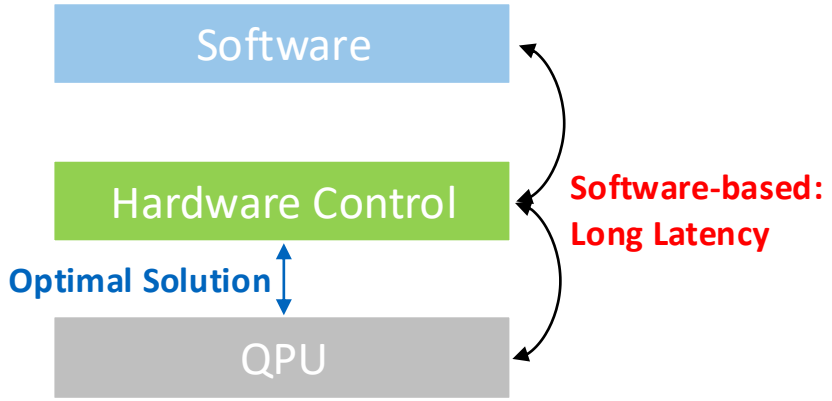
Mid-circuit Measurement

- Current architecture mostly requires simultaneous readout of all qubits
- Hindering the applicability in mid-circuit measurement scenarios
  - Individual readout is required



Migration to FPGAs

- Software-based methods -> Long process & communication latency
- Existing designs are too heavy to scale to large qubit numbers on the hardware



NEED! Lightweight FNN to support independent qubit readout

Proposal

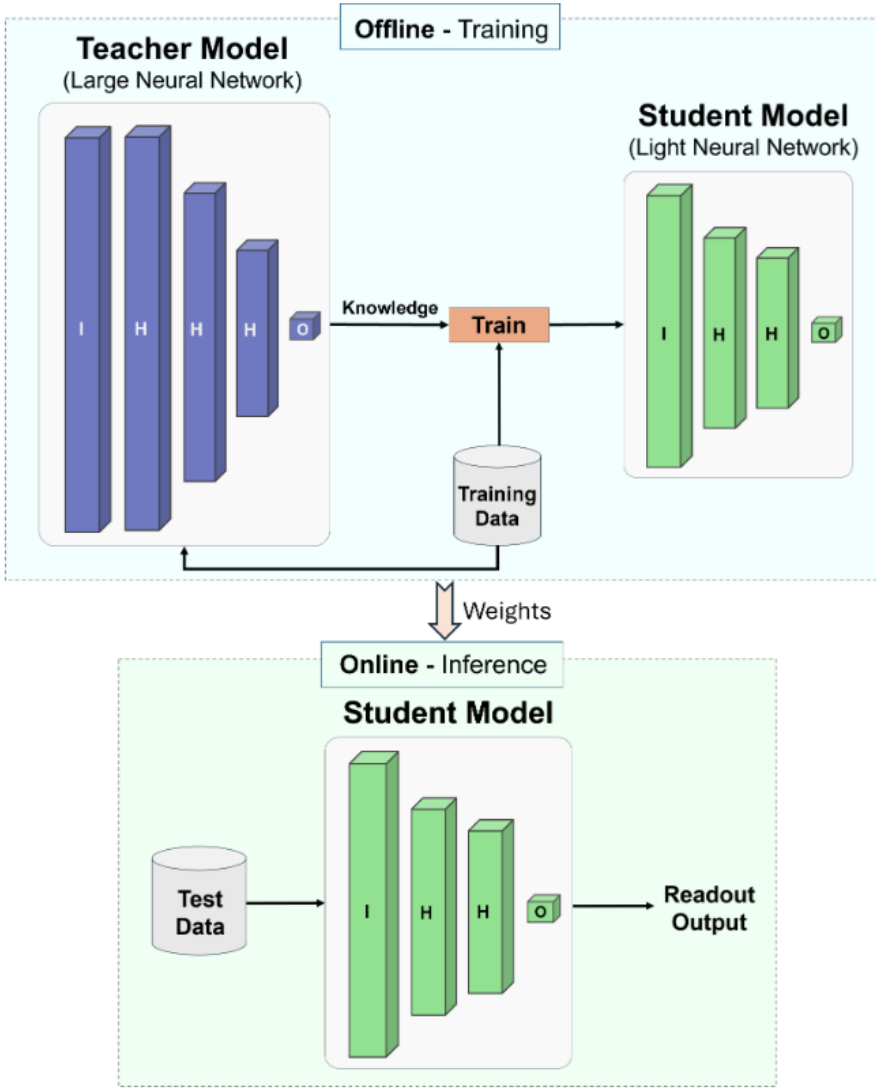
- Using Knowledge-Distillation to Reduce the FNN Size

Independent Readout

- Each qubit has an individual readout discriminator

Distilled Lightweight FNN

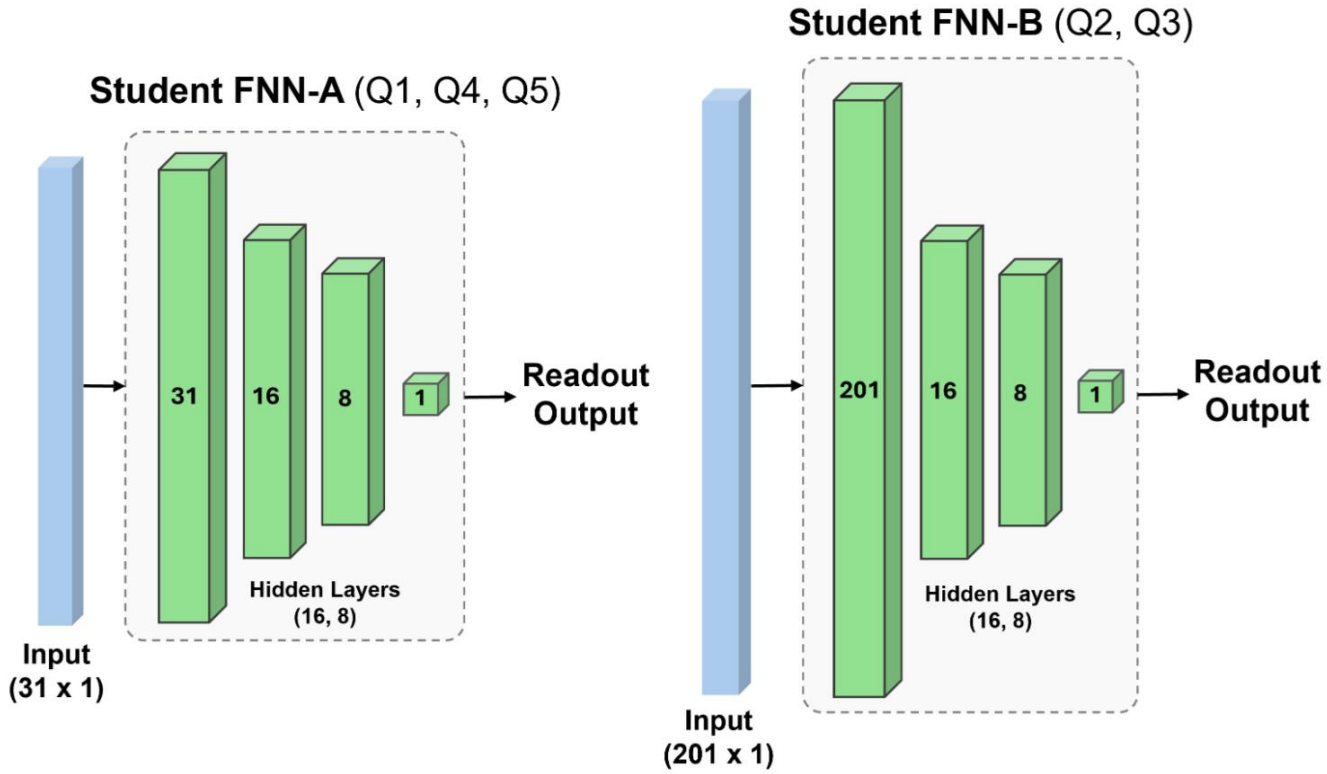
- Offline – Training
  - Utilize knowledge distillation to distill the information and train a small size network
- Online – FPGA inference
  - Deploy small (lightweight) model on the FPGA



Lightweight Neural Network

Specialized Configuration

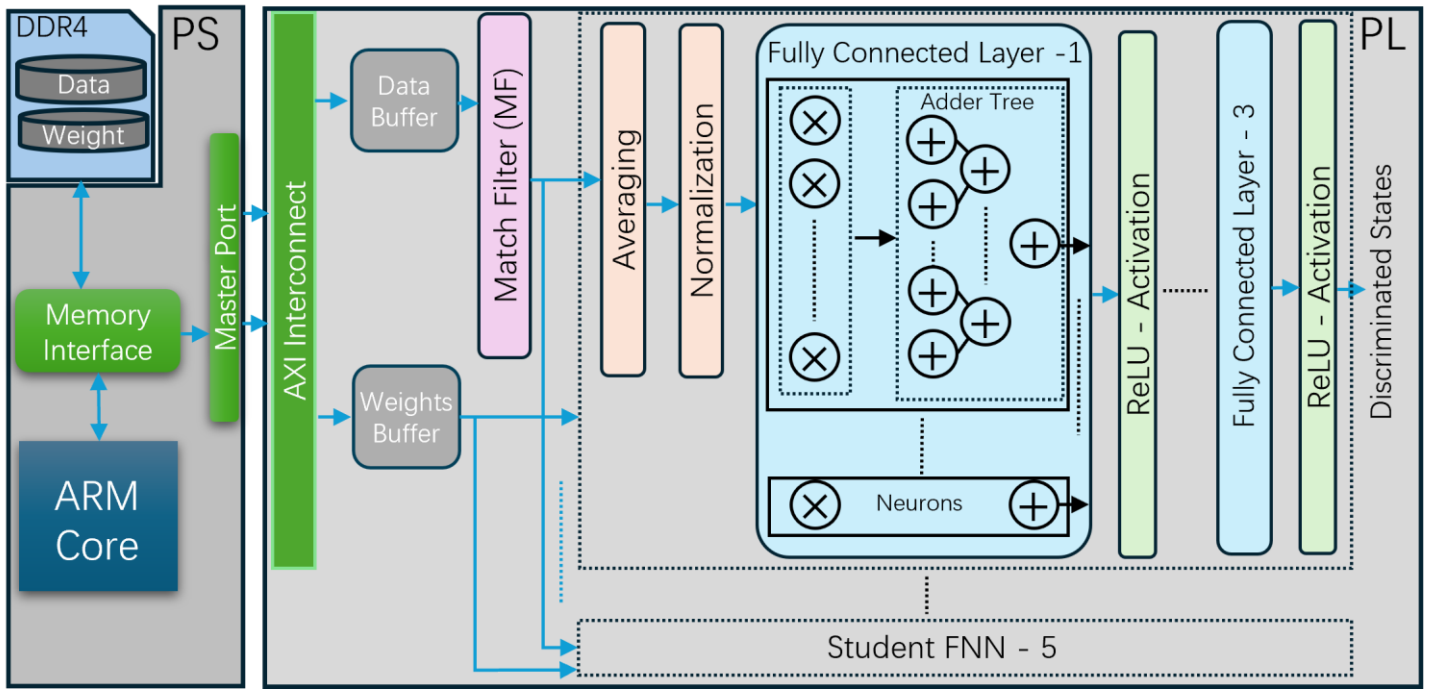
- FNN-A
  - Target qubits 1, 4, and 5 experience less crosstalk during measurement
  - Average every 32 samples plus the match filter parameter
- FNN-B
  - Target qubits 2 and 3, which suffer significantly from crosstalk, exhibit lower accuracy when evaluated by the teacher model
  - Average every 5 samples plus the match filter parameter



FPGA Implementation

Specifications

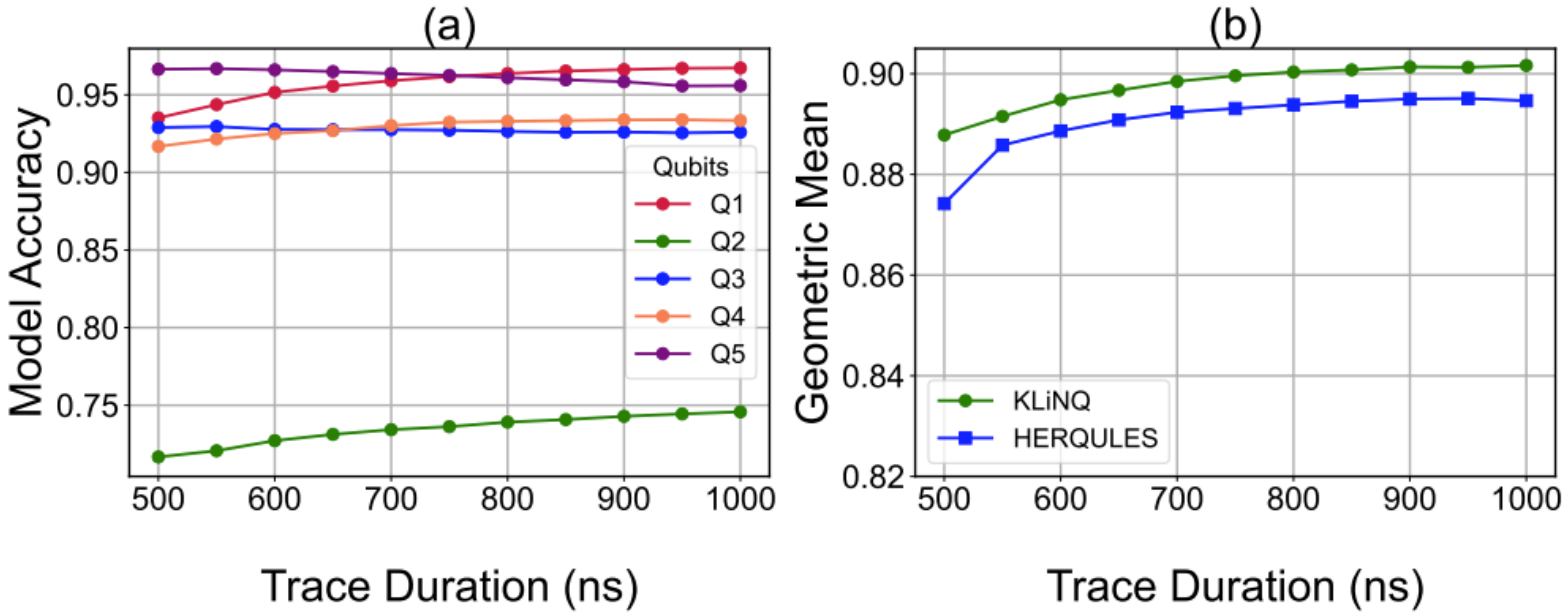
- SoC-based implementation
  - DDR: Storing raw quantum data and pre-trained weights
  - PS: Data exchange interface
  - PL: Customized FNN Accelerator
- Design Highlights
  - Individual FNN architectures for each qubits, but share MF module
  - HW-SW Codesign optimization within Averaging and Normalization Modules
  - Four-stage pipelined multiplication (time-multiplexed DSP resources) plus Adder Trees
  - Quantization of data and network weights



Experimental Results

Readout Accuracy (Fidelity)

- Calculated by geometric mean
- Robustness across the trace duration
- Geometry mean of 0.906 when measurement trace equals 1us
- Perform much better than the best candidate



Experimental Results

Readout Accuracy (Fidelity)

- Calculated by geometric mean
- Robustness across the trace duration
- Geometry mean of 0.906 when measurement trace equals 1us
- Perform much better than the best candidate

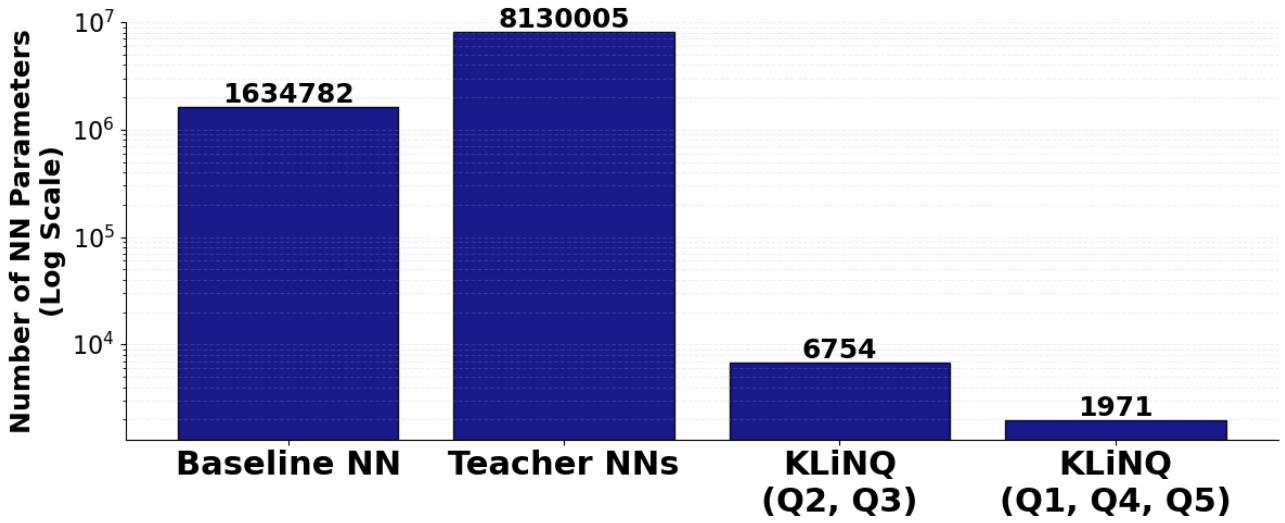
Design	Duration	Qubit 1	Qubit 2	Qubit 3	Qubit 4	Qubit 5	F_{5Q}
KLiNQ NNs	1μs	0.968	0.748	0.929	0.934	0.959	0.904
	950ns	0.967	0.744	0.925	0.934	0.956	0.901
	750ns	0.962	0.736	0.927	0.932	0.963	0.900
	550ns	0.944	0.720	0.930	0.921	0.967	0.891
	500ns	0.935	0.717	0.929	0.917	0.966	0.887

Design	Qubit 1	Qubit 2	Qubit 3	Qubit 4	Qubit 5	F_{5Q}	F_{4Q}
Baseline FNN[1]	0.969	0.748	0.940	0.946	0.970	0.910	0.956
HERQULES[2]	0.965	0.730	0.908	0.934	0.953	0.893	0.940
KLiNQ	0.968	0.748	0.929	0.934	0.959	0.904	0.947

Experimental Results

Network Compression

- Around 99% network compression rate compared to the teacher FNN
- Around 98% network compression rate compared to the baseline model



Experimental Results

FPGA-based Results

- 32 ns latency for the trace processing
- Acceptable hardware result utilization

Components	LUT	FF	DSP	Latency (ns)
Shared Resources				
MF	27180 (6.39%)	24052 (2.83%)	375 (8.78%)	11
Per-Qubit Resources: Qubits 1, 4, 5				
AVG&NORM	17770 (4.2%)	11415 (1.35%)	0 (0%)	9
Network	8840 (2.08%)	6020 (0.71%)	55 (1.28%)	12
Per-Qubit Resources: Qubits 2, 3				
AVG&NORM	19600 (5%)	17500 (2%)	0 (0%)	6
Network	25882 (8.44%)	23172 (2.72%)	226 (5.29%)	15

Conclusion & Outlook

Conclusion

- We proposed a method leveraging knowledge distillation to train lightweight neural networks optimized for FPGA implementation.
- Independent readout for mid-circuit measurement
- 91% readout accuracy
- Only 32ns readout latency

Federal Ministry of Education and Research

Bayerisches Staatsministerium für Wissenschaft und Kunst

Future work:

- Explore complex machine learning models
- Optimize hardware design to increase scalability

Contact:

- xiaorang.guo@tum.de
- https://www.ce.cit.tum.de/caps/mitarbeiter/xiaorang-guo/